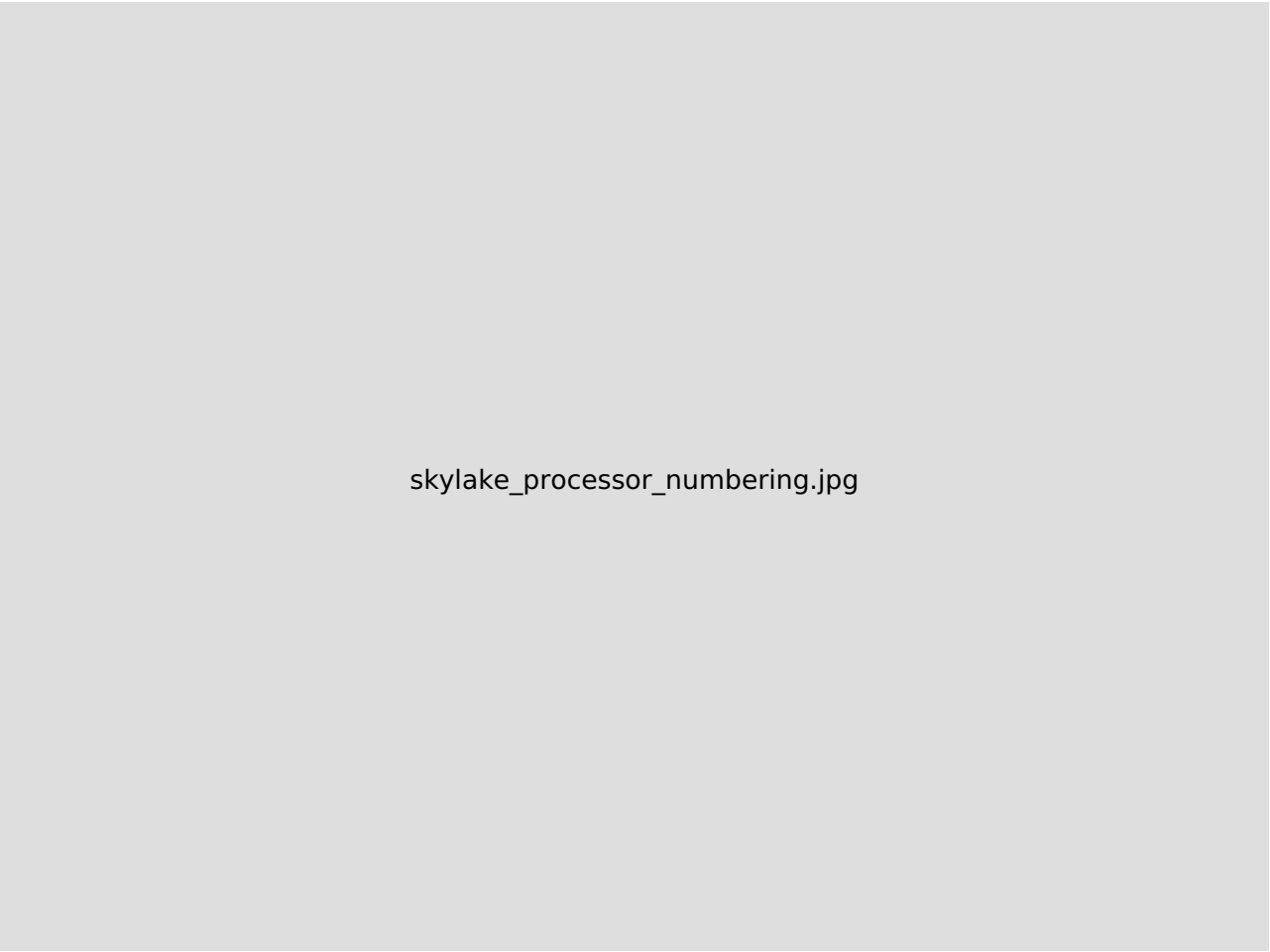
Skylake Processors



The Intel Skylake processor incorporated into the Electra cluster is the 20-core Xeon Gold 6148 model with a base clock speed of 2.4 GHz. Skylake is a microarchitecture redesign using the same 14-nanometer (nm) manufacturing process technology as Broadwell, with multiple new features and enhancements in the on-chip and intersocket interconnects, memory, cache, and CPU.

Each Electra Skylake node makes use of the 100 Gbits/s four-lane Enhanced Data Rate (4X EDR) technology to connect to the rest of the Pleiades/Electra InfiniBand network.

Each small square in the diagram above represents a combination of a physical core and a L3 cache slice. For each physical core, there are two logical core labelings obtained from the "processor" entry of the **`cat /proc/cpuinfo`** output.

On-chip and Inter-socket Interconnects

In previous generations of Intel Xeon processors (such as Sandy Bridge, Ivy Bridge, Haswell, and Broadwell), the cores, L3 cache, memory controller, I/O controller and intersocket interconnect ports are connected with a on-chip ring architecture. This has the drawback of increased latency and bandwidth constraints when the number of cores per socket increases. To mitigate this issue, Skylake utilizes a mesh architecture that encompasses an array of vertical and horizontal paths, allowing communication from one core to another through a shortest path. Furthermore, each core and L3 slice has a combined Caching and Home Agent (CHA) that handles address

mapping and information routing, and provides scalability of resources across the mesh.

The two sockets are connected with two Intel Ultra Path Interconnect (UPI) links, which deliver increased bandwidth and performance over the Intel Quick Path Interconnect (QPI) that was used in previous generations of Intel Xeon processors. The UPI runs at a speed of 10.4 gigatransfers per second (GT/s). Each link contains separate lanes for the two directions. The total full-duplex bandwidth (2 links x 2 directions) is 41.6 gigabytes per second (GB/s).

Note: With full-duplex communication between two components, both ends can transmit and receive information between each other simultaneously. With half-duplex communication, the transmission and reception must happen alternatively.

Memory Subsystem

There are two sub-NUMA clusters in each socket, creating two localization domains. Each domain contains one memory controller and ten L3 cache slices. Processes observe lower L3 and memory latency if they access data mapped to the L3 or memory in the same domain where the processes are running, compared with outside of the domain.

There are three memory channels per sub-NUMA cluster. Each channel can be connected with up to two memory DIMMs. For the Electra Skylake configuration, there is one 16-GB dual rank DDR4 DIMM with error correcting code (ECC) support per channel. In total, the amount of memory is 48 GB per sub-NUMA cluster, 96 GB per socket, and 192 GB per node.

The activation or deactivation of the sub-NUMA domains is controlled in the Basic Input/Output System (BIOS) firmware. Currently, they are deactivated for the Electra Skylake nodes.

The speed of each memory channel is 2,666 MHz. One 8-byte read or write can take place per cycle per channel. With a total of 6 memory channels, the total half-duplex memory bandwidth is approximately 128 GB/s per socket.

Cache Hierarchy

The cache hierarchy of Skylake is as follows:

- L1 instruction cache: 32 KB, private to each core; 64 B/line; 8-way
- L1 data cache: 32 KB, private to each core; 64 B/line; 8-way; fastest latency: 4 cycles
- L2 cache: 1 MB, private to each core; ; 64 B/line; 16-way; fastest latency: 12 cycles
- L3 cache: shared non-inclusive 1.375 MB/core; total of 27.5 MB, shared by 20 cores in each socket; fully associative; fastest latency: 44 cycles

In Broadwell, the L2 cache is 256 KB per core and the L3 cache is a shared inclusive cache with 2.5 MB per core. In Skylake, the cache hierarchy has changed to provide a larger L2 cache of 1 MB per core and a smaller shared non-inclusive 1.375 MB L3 cache per core.

Note: An inclusive L3 cache guarantees that every block that exists in the L2 cache also exists in the L3 cache. A non-inclusive L3 cache does not guarantee this.

A larger L2 cache increases the hit rate into the L2 cache, resulting in lower effective memory latency and lower demand on the mesh interconnect and L3 cache.

If the processor has a miss on all the levels of the cache, it fetches the line from memory and puts it directly into the L2 cache of the requesting core, rather than putting a copy into both the

L2 and L3 caches, as is done on Broadwell. When the cache line is evicted from the L2 cache, it is placed into L3 if it is expected to be reused.

Due to the non-inclusive nature of the L3 cache, the absence of a cache line in L3 does not indicate that the line is absent in private caches of any of the cores. Therefore, a snoop filter is used to keep track of the location of cache lines in the L1 or L2 caches of cores when a cache line is not allocated in L3. On the previous-generation processors, the shared L3 itself takes care of this task.

Processor

Like the Broadwell processors, Skylake supports single instruction, multiple data (SIMD) instruction sets, including several generations of Streaming SIMD Extensions (SSE, SSE2, SSE3, Supplemental SSE3, SSE4.1 and SSE4.2), and Advanced Vector Extensions (AVX and AVX2). In addition, it also includes MPX (Memory Protection Extensions), Intel SGX (Software Guard Extensions) and Advanced Vector Extensions 512 (AVX-512).

AVX-512 was originally introduced with the Intel Xeon Phi processors. Of the multiple AVX-512 instruction groups, Skylake comes with the AVX512F, AVX512CD, AVX512BW, AVX512DQ groups and a new AVX512VL feature.

With AVX-512, programs can pack eight double precision or 16 single precision floating-point numbers, or eight 64-bit integers, or 16 32-bit integers within the 512-bit vectors. With 512-bit floating-point vector registers and two floating-point functional units, each capable of Fused Multiply-Add (FMA), a Skylake core can deliver 32 floating-point operations per cycleâ double the number of operations of a Haswell/Broadwell core, or quadruple that of a Sandy Bridge/Ivy Bridge core can deliver.

Unlike SSE and AVX, which cannot be mixed without performance penalties, the mixing of AVX and Intel AVX-512 instructions is supported without penalty. AVX registers YMM0-YMM15 map into the Intel AVX-512 registers ZMM0-ZMM15, very much like SSE registers map into AVX registers. Therefore, in processors with Intel AVX-512 support, AVX and AVX2 instructions operate on the lower 128 or 256 bits of the first 16 ZMM registers.

Intel AVX-512 optimizations are included in Intel compiler version 16.0 and later versions. For Skylake-specific optimizations, use **`-xCORE-AVX512`**. With Intel compiler 2018 versions, using the optimization flag **`-qopt-zmm-usage=high -xCORE-AVX512`** may benefit floating-point heavy applications running on Skylake.

TIP: If you want a single executable that will run on any of the Electra or Pleiades processor types, with suitable optimization to be determined at run time, you can compile your application using the option **`-O3 -axCORE-AVX512 -xSSE4.2`**.

Hyperthreading

Hyperthreading is turned ON.

Turbo Boost

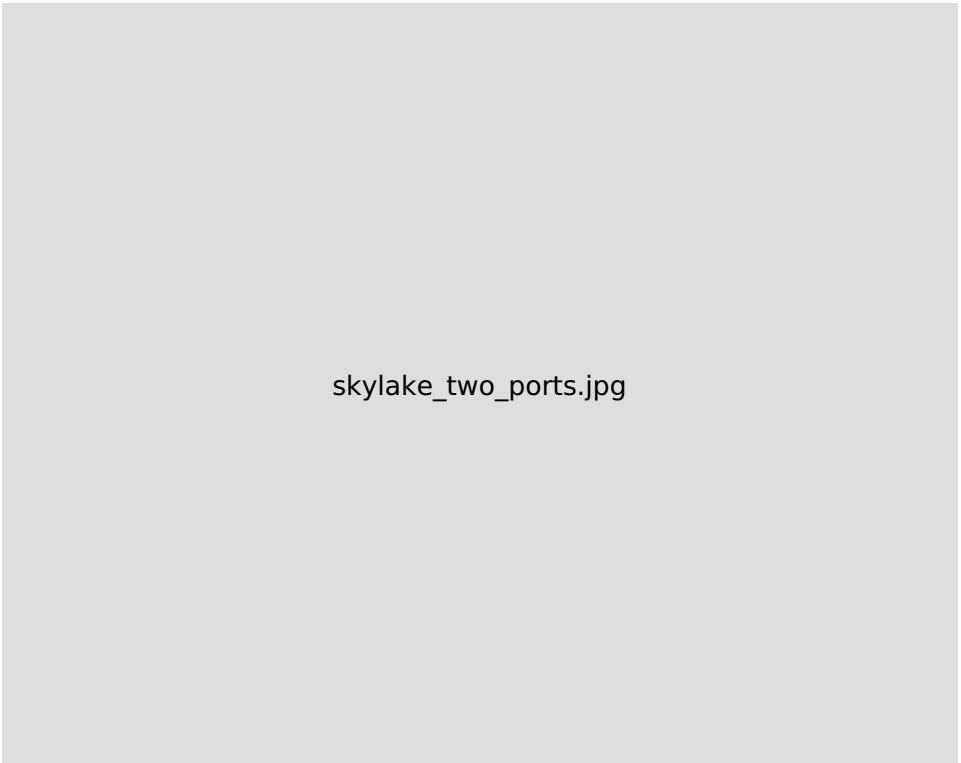
Turbo Boost is turned ON. Maximum Turbo Frequency is 3.70 GHz.

Inter-Node Network

Like the network subsystem of the Pleiades Haswell and Broadwell nodes, each Skylake node uses two PCI Express interfaces (one from each socket) to connect to the Pleiades and Electra Infiniband fabric. The use of PCIe 3.0 16-lane (x16) links enables a maximum bandwidth of 15.75 GB/s for each direction.

One enhancement in the NAS Skylake inter-node network is that it makes use of the 100 Gbits/s Enhanced Data Rate technology instead of the 56 Gbits/s Fourteen Data Rate technology used for the NAS Pleiades nodes (Sandy Bridge, Ivy Bridge, Haswell and Broadwell nodes). As shown below, one PCIe is connected to the ib0 Infiniband fabric via a single-port, four-lane, Enhanced Data Rate (4X EDR) host channel adapter (HCA), in a dual single-port EDR IB mezzanine card, with an effective bandwidth of 100 Gbits/s (that is 12.5 GB/s) for each direction. The other PCIe is connected to the ib1 fabric via another single-port 4x EDR HCA in the same mezzanine card.

Note: For nodes labeled with r[x]i[x]n[9-17,27-35], socket 0 is connected to ib0 and socket 1 is connected to ib1. For nodes labeled with r[x]i[x]n[0-8,18-26], socket 0 is connected to ib1 and socket 1 is connected to ib0.



There are 2,304 Skylake nodes in the Electra cluster. They are partitioned as follows:

- Eight E-cells (288 nodes per E-cell)
- Two compute racks and one cooling rack per E-cell (144 nodes per rack)
- Four individual rack units (IRUs) per compute rack (36 nodes per IRU)
- Nine compute blade (ICE XA SAXON blade) per IRU (4 nodes per blade)

The connection of the compute nodes relies on the use of four premium IB EDR switch blades per IRU. Each switch blade has two 36-port ConnectX-5 ASICs with 9 ports from each ASIC connecting to compute nodes and 18 ports from each ASIC for connecting to external targets. The connection of the 2,304 Skylake nodes forms an 8-dimensional hypercube. The topology of Electra (including both the Skylake and Broadwell nodes) is a 9-dimensional hypercube.

MPT 2.15 and older versions do not support the ConnectX-5 HCAs. To avoid jobs failing when using MPT 2.15 and older versions, the environment variables **`MPI_IB_XRC`** and **`MPI_XPMEM_ENABLED`**

have been disabled for jobs running on Skylake. If your MPI applications perform significant MPI collective operations and rely on having these two variables enabled to get good performance, use MPT 2.16 or the forthcoming newer versions.

References

- Intel Xeon Processor Scalable Family Technical Overview
- Intel 64 and IA-32 Architectures Optimization Reference Manual
- HPE SGI 8600 QuickSpecs

Article ID: 550
Last updated: 13 May, 2021
Revision: 12
Systems Reference -> Electra -> Skylake Processors
https://www.nas.nasa.gov/hecc/support/kb/entry/550/